DOCUMENT RESUME

ED 334 271 TM 016 919

AUTHOR Lam, Tit-Loong; Foong, Yoke-Yeen
TITLE Development and Evaluation of Hierarchical Testlets in Two-Stage Tests Using Integer Linear Programming.
PUB DATE Apr 91
NOTE 21p.; Paper presented at the Annual Meeting of the American Educational Research Association (Chicago, IL, April 3-7, 1991).
PUB TYPE Reports - Research/Technical (143) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *Adaptive Testing; Algorithms; *Computer Assisted Testing; Computer Simulation; Evaluation Methods; Item Response Theory; *Linear Programing; Mathematical Models; *Test Construction; Test Interpretation
IDENTIFIERS *Integer Programming; Testlets; *Two Stage Testing

ABSTRACT

This simulation study involved the design of two two-stage tests in which the routing tests and the second-stage measurement testlets took the form of a multidimensional knapsack problem with prespecified target informations and constraints to be enumerated using the algorithm of E. Balas. Two conventional tests of similar length to the two-stage tests were developed by random selection of items from the same item pool. Item responses of 1,000 simulees on 240 items were generated using the computer program DATAGEN based on the three-parameter item response model. Analysis of the degree in which the true abilities were recovered showed that the inaccuracies and root square mean errors for the two-stage designs were lower than those for their conventional counterparts. Fidelity coefficients were also higher and compared well with the results of previous adaptive test procedures. Three tables and three graphs accompany the discussion. (Author/SLD)

**********
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
**********

ED334271

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☑ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

TIT-LOONG LAM

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

# DEVELOPMENT AND EVALUATION OF HIERARCHICAL TESTLETS IN TWO-STAGE TESTS USING INTEGER LINEAR PROGRAMMING[1]

Tit-Loong Lam[2]
Yoke-Yeen Foong[3]

University of Massachusetts at Amherst

Paper presented at the 1991 Annual Meeting of the American Educational Research Association, Chicago.

---

[1]The authors wish to thank Professor Ron Hambleton and Professor S. Swaminathan for their comments and suggestions in the preparation of this paper.

[2,3]Both authors on study leave from the National Institute of Education, Singapore.



TM016919

BEST COPY AVAILABLE

DEVELOPMENT AND EVALUATION OF HIERARCHICAL TESTLETS IN TWO-STAGE TESTS USING INTEGER LINEAR PROGRAMMING


## Abstract

This simulation study involved the design of 2 two-stage tests in which the routing tests and the second stage measurement testlets took the form of a multidimensional 'knapsack problem' with prespecified target informations and constraints to be enumerated using Balas' algorithm. Two conventional tests of similar length to the two-stage tests were developed by random selection of items from the same item pool. Analysis of the degree in which the true abilities were recovered showed that the Inaccuracies and the RMSEs for the two-stage designs were lower than their conventional counterparts. Fidelity coefficients were also higher and compared well with the results of previous adaptive test procedures.



*Index terms: adaptive tests, two-stage tests, integer linear programming, test development, test information*

# DEVELOPMENT AND EVALUATION OF HIERARCHICAL TESTLETS IN TWO-STAGE TESTS USING INTEGER LINEAR PROGRAMMING

Tit-Loong Lam, Yoke-Yeen Foong
University of Massachusetts at Amherst

A number of concerns in adaptive testing have been highlighted by various researchers. Among some of these concerns are content bias (Wise and Plake, 1989), context effects and item exposure (Wainer and Kiely, 1987).

In an attempt to address these issues, Wainer and Kiely (1987) proposed the testlet model in which the examinee is given a fixed number of predetermined paths in a pyramidal item selection procedure. Kingsbury and Zara (1989) however, criticized the use of the testlet model on the grounds that this will reduce measurement accuracy because of the 'weak' prestructured selection strategy. Furthermore, the use of pyramidal selection strategy is rather inefficient as it requires a rather large item pool. Kingsbury and Zara (1989) proposed a constrained version of CAT whereby a number of components are built into the selection algorithm. These include a content-balancing algorithm whereby, items selectively take into consideration, the test specification of the test developer.

The rationale for the administration of hierarchical testlets instead of single items in CAT is that it has the advantage of limiting context effects and item exposure. The purpose of this study is to improve upon the pyramidal testlet design of Wainer and

Kiely (1987) by using a two-stage format in administering hierarchical testlets and making use of integer linear programming in the test design process.

The test design is treated as a 'multidimensional Knapsack problem', one of the more popular problem type in Optimization Theory. Researchers who made use of linear programming for test designs (e.g. Theunissen, 1985, 1986) evaluated their algorithms and models by comparing the obtained test information curves with the target information as well as obtaining item statistics and the time taken to enumerate the optimization problem. This study e04extends the evaluation of test designs further by determining the accuracy in which the true abilities of the simulees are recovered by the test procedure.

## Method

Item responses of 1000 simulees on 240 items were generated using the computer program, DATAGEN (Hambleton and Rovinelli, 1973) based on the 3 parameter IRT model with the following distribution: $b_i \sim U(-3.0,3.0)$, $a_i \sim U(0.5,2.0)$, $c_i \sim U(0.0,0.20)$ and $\theta_e \sim N(0,1)$. The characteristic of the item bank is based on the suggestions of Urry (1977) where a uniform item bank with highly discriminating items is most desirable.

Based on suggestions for the development of two-stage tests by Theunissen (1985, 1986) and Adema (1990), the test information desired was specified for five $\theta$ points for the routing test and

the test design takes the form of a multidimensional KP problem as follows:

Minimize the objective function:

$$n = x_1 + x_2 + \ldots . x_{n1}$$

Subject to the constraints:

$$I(\theta_j,1)x_1 + I(\theta_j,2)x_2 + \ldots . I(\theta_j,n1)x_{n1} \geq b_j$$

so that $x_i = 0$ or 1, $i = 1,2,\ldots n_1$, $j = 1,2,\ldots 5$

where $n_1$ is the size of the item bank stratum.

The item bank was divided into two strata of 120 items each for the purpose of developing the routing test and second stage tests separately. This is to ensure an even distribution of items between two pools in which the routing test and the second stage tests were developed.

$b_i$ represents the target information for five specified ability levels: -1.75,-1.00,0.00,1.00,1.75. Enumeration of this problem involves optimal selection of the minimum number of items that will fit the target information. For this study, two uniform targets were defined in the range $-1.75 \leq \theta \leq 1.75$ with $I(\theta) = 8$ and $I(\theta) = 4$. This region includes more than 92% of the population of simulees. Two forms of the two-stage test were therefore developed as a result of setting two routing test target informations.

The design of the second-stage test was such that the items were selected to give maximum information at the specified ability segments given a certain test length:

Maximize: $I(\theta_j,1)x_1 + I(\theta_j,2)x_2 \ldots + I(\theta_j,n2)x_{n2}$ (3)

Subject to the constraints:

$x_1 + x_2 \ldots + x_{n2} = n_t$ , $x_i = 0$ or 1, $i = 1,2, \ldots n_2$ (4)

and $j = 1,2, \ldots 10$ for the 10 ability levels

where $n_t$ is the length of the desired 2nd-stage test, set at 15 for this study and $n_2$ is the size of the item bank stratum meant for developing the second-stage tests. This second-stage test involved ten clusters of items to measure 10 equally spaced ability segments to cover the whole trait range. Each group of items will have maximum information at the mid-point of the corresponding ability range. Maximum information levels will therefore be at ability levels, from -2.25 to +2.25, inclusive at intervals of 0.5.

Balas' algorithm (Balas, 1965) was used in the solution of this test construction model. This algorithm involves a combinatorial approach to solving linear programs with zero-one variables. This method is actually a tree-search algorithm that uses information generated in the search to exclude portions of the tree under consideration. A simplified version of the algorithm together with a FORTRAN listing is given by McMillan (1970). The paths in the algorithm and the enumerations involved could be followed easily using any statistical package that has a macro language component.

The software, STATA (Computer Resource Center, 1990) was used for this purpose by making use of its programming language. The resulting response strings were created to serve as inputs into

the MicroCAT (Assessment Systems Corporation, 1987) environment for conventional scoring.

The simulee responses were scored on the items selected by the routing test and their abilitiy levels estimated by the Bayesian-modal procedure. Based on these initial ability estimates, the simulees were divided into 10 ability groups with the range: $-2.5 \leq \theta \leq +2.5$ with an interval of 0.5. Simulees were routed to their respective 2nd-stage measurement testlets where their responses were scored by maximum-likelihood procedure. Each simulee would therefore have two ability estimates - one from the routing test and the other from the testlet. The two ability estimates were averaged after weighting them inversely according to their variances (Lord, 1980). The overall ability estimate of each simulee is given as:

$$\theta' = (\theta_1 \mathrm{Var}(\theta_2) + \theta_2 \mathrm{Var}(\theta_1))/(\mathrm{Var}(\theta_1) + \mathrm{Var}(\theta_2))$$

Two conventional tests were developed by randomly selecting items from the same item pool, the test lengths of which would be the same as the lengths of the two-stage tests. The conventional tests are expected to be broad-range tests since they are developed from the item pool which has uniformly distributed and highly discriminating items. Based on the response strings of the same simulees on the item pool, the conventional tests were scored by maximum likelihood procedure to correspond to the same procedure used for scoring the testlets.

Apart from computing the fidelity coefficients (Weiss & McBride, 1984), two evaluative indices were used in this study. For

each ability group, the Inaccuracy is computed as:

$$\text{Inacc}(\theta) = \Sigma \mid (\theta_e - \theta_t) \mid /n$$

where: n is the number of examinees in the ability grouping, $\theta_e$ is the estimated ability and $\theta_t$ is the true ability.

The root mean square error is computed as:

$$RMSE(\theta) = \sqrt{\frac{\sum (\theta_e - \theta_t)^2}{n}}$$

This index gives more weight to larger differences between estimated and true abilities. Small Inaccuracies or RMSEs will imply estimates that are closer to the true abilities and hence, a greater level of accuracy in ability estimation for that ability grouping.

## Results

Solution of the KP problem resulted in minimum routing test lengths of 18 and 10 items for the uniform target informations set at 8 and 4 respectively. Tables 1 and 2 show the descriptive statistics of parameters for items of the routing tests and the conventional tests. Although the trait range covered by the items was restricted by the target information, it can be seen that the information curves of the routing tests were flatter and smoother compared to those of the conventional tests (see Figs 1 and 3). This is the effect of proper test design and enumeration technique. For the two routing tests, extreme items with b values greater than

2.00 or less than -2.10 were not selected. On the other hand, because of no restrictions imposed on the development of the conventional tests, the items covered a broader range of difficulty levels and hence, the associated ability levels.

---

Insert Tables 1, 2 and 3 about here

---

Table 3 shows the distribution of the item parameters of the 10 testlets. As expected, the increasing mean difficulties of the series of 15-item testlets, each having high item discrimination is indicative of the hierarchical nature of the testlets. The testlets are actually a series of peaked conventional tests as shown by the plots of the testlet information (see Figure 2). If the simulees are correctly routed to the second-stage testlets, measurement precison will be optimum for that trait range. The uniform distribution of highly discriminating testlets will result in equiprecise measurement along a wide range of the ability continuum.

---

Insert Figs 1, 2 and 3 about here

---

Fig 4 shows a comparison of the inaccuracies of the two-stage tests compared with those of the conventional tests. It can be seen that both two-stage tests showed lower inaccuracies in the whole range of the ability levels compared to their conventional

counterparts. This shows that the lower inaccuracies resulted from correct routing of the simulees to their respective hierarchical testlets. The pattern of inaccuracies between the two forms of the two-stage tests was almost the same. The two conventional tests however, although covering a wider ability spectrum showed greater inaccuracies at the ends of the trait level. This is an indication of the bandwidth-fidelity problem where a greater measurement precision is found for tests developed to measure precisely at a restricted range at the expense of covering a broader range of the trait because of restrictions on the number of items.

---

Insert Fig 4 about here

---

Examination of the RMSE of the ability estimates (see Fig 5) shows that the variability of errors resulting from both forms of the two-stage tests is consistently lower than those of their conventional counterparts. However, the 25-item two-stage test shows greater variability at the range below 0.0 The fidelity coefficients are 0.98 and 0.97 for the 33-item and 25-item two-stage tests respectively compared to 0.94 and 0.93 for their conventional counterparts.

---

Insert Fig 5 about here

---

## Discussion

The results showed the efficiency of well designed two-stage testing compared to conventional testing. The purpose of two-stage testing is not to minimize the number of items to be administered, but to enhance the precision of estimation of examinee ability (Lord, 1980). In a sense, two-stage testing is adaptive since in this case, the simulees were given a combination of testlets suited to their ability level. Two advantages can be seen from this test model. Firstly, the test developer is in full control of the whole testing process, beginning with test design. It is possible for testlets to be stored in the database as sets of hierarchical testlets instead of single items. The test developer can have the liberty to alter the target information of the test and to change the test length. Hambleton and Swaminathan (1985) suggested that knowledge of the examinees can be used for test construction purposes. As such, the target information can be shifted based on prior knowledge of the examinee ability distribution. Secondly, there is a possible reduction of context effects because the testlets are administerd instead of individual items. In a normal adaptive testing system, every examinee is administered a different set of items giving rise to possible context effects due to the different location of the test items. Administration of hierarchical testlets instead of individual items means that the same set of items are administered to a larger group of examinees,

limiting context effects to sets of testlets instead of multiple sets of items.

The results of this study showed that the RMSEs and Inaccuracies of the two-stage tests are small and their patterns compare well with the results obtained by previous studies on adaptive testing using either maximum information item selection or minimum standard error item criterion (Weiss, 1982; De Ayala, Dodd & Koch, 1990). The good fidelity coefficients of the two-stage tests are equivalent to using a posterior variance criterion of 0.0625 (or SE of estimate of 0.25) as a termination rule in Bayesian sequential adaptive testing (McBride, 1977) which will result in a fidelity coefficient of greater than 0.968. The study showed that even though the uniform target information was set at a low level of 4 to design the routing test, the simulees were accurately routed to the appropriate testlets.

## Conclusion

The study showed the effectiveness of using appropriately designed two-tests in trait estimation to limit context effects and to allow for control of test design. Better results in recovering the true abilities were obtained compared to conventional tests where no test design was applied.

This study was based on an ideal item pool where the highly discriminating items are uniformly distributed. Further, since the parameters were generated, they are assumed to be estimated without

errors. A possible extension of this study is to use a realistic item pool based on life data. Stratification of the item pool can then be based on content subdomains where the routing test and second stage testlets represent separate content categories. Further study can be done with more content subdomains as in multistage testing.

# Bibliography

Adema, J.J. (1990). The construction of customized two-stage tests. Journal of Educational Measurement, 27, 241-253.

Assessment Systems Corporation (1987). User's manual for the MicroCAT. St Paul, Minnesota: Author.

Balas, E. (1965). An additive algorithm for solving linear programs with zero-one variables. Operations Research, 13, 517-546.

Computer Resource Center (1990). User's reference manual for STATA. Santa Monica, CA: Author.

De Ayala, R.J., Dodd, B.G. & Koch, W.R. (1990). A simulation and comparison of flexilevel and Bayesian computerized adaptive testing. Journal of Educational Measurement, 27, 227-239.

Hambleton, R.k. & Rovinelli, R. (1973). A Fortran IV program for generating examinee response data from logistic test models. Behavioural Science, 17, 73-74.

Hambleton, R.K. & Swaminathan, H. (1985). Item response theory: Principles and applications (pp.225-253). Boston: Kluwer Nijhoff.

Kingsbury, G.G. & Zara, A.R. (1989). Procedures for selecting items for computerized adaptive tests. Applied Measurement in Education, 2, 359-375.

Lord, F.M. (1980). Applications of item response theory to practical testing problems. Hillsdale, NJ: Erlbaum.

McBride, J.R. (1977). Some properties of a Bayesian adaptive ability testing strategy. Applied Psychological Measurement, 1, 121-140.

McMillan, Claude Jr. (1970). Mathematical programming: an application to the design and application of optimal decision machines. New York: John Wiley.

Theunissen, T.J.J.M. (1985). Binary programming and test design. Psychometrika, 50, 411-420.

Theunissen, T.J.J.M. (1986). Optimization algorithms in test design. Applied Psychological Measurement, 10, 381-389.

Urry, V.W. (1977). Tailored testing: A successful application of latent trait theory. Journal of Educational Measurement, 14, 182-196.

Wainer, H. & Kiely, G.L. (1987). Item cluseters and computerized adaptive testing: a case for testlets. Journal of Educational Measurement, 24, 185-201.

Weiss, D.J. (1982). Improving measurement quality and efficiency with adaptive testing. Applied Psychological Measurement, 6, 473-492.

Weiss, D.J. & McBride, J.R. (1984). Bias and information of Bayesian adaptive testing. Applied Psychological Measurement, 8, 273-285.

Wise, S.L. & Plake, B.S. (1989). Research on the effects of administering tests vis computers. Educational Measurement, Issues and Practice, 8, 5-10.

Table 1

Descriptive statistics of item parameters of routing tests

| Routing Test 1 (Target infor=8) | Mean | SD | Min | Max |
|---|---|---|---|---|
| a | 1.83 | 0.09 | 1.67 | 1.94 |
| b | -0.03 | 1.41 | -2.07 | 1.80 |
| c | 0.08 | 0.06 | 0.01 | 0.19 |
| **Routing Test 2 (Target infor=4)** | | | | |
| a | 1.75 | 0.16 | 1.46 | 1.94 |
| b | 0.01 | 1.33 | -2.07 | 1.90 |
| c | 0.08 | 0.05 | 0.02 | 0.18 |

Table 2

Descriptive statistics of item parameters of conventional tests

| Test Length = 25 | Mean | SD | Min | Max |
|---|---|---|---|---|
| a | 1.36 | 0.35 | 0.72 | 1.93 |
| b | 0.18 | 1.76 | -2.49 | 2.98 |
| c | 0.12 | 0.05 | 0.03 | 0.19 |
| **Test Length = 33** | | | | |
| a | 1.29 | 0.35 | 0.59 | 1.88 |
| b | 0.14 | 1.69 | -2.87 | 2.72 |
| c | 0.10 | 0.05 | 0.00 | 0.19 |

Table 3

Descriptive statistics of item parameters of 15-item testlets

| | Testlet | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
| Mean (a) | 1.63 | 1.66 | 1.69 | 1.63 | 1.62 | 1.65 | 1.62 | 1.56 | 1.46 | 1.50 |
| SD (a) | 0.26 | 0.24 | 0.23 | 0.25 | 0.20 | 0.16 | 0.27 | 0.34 | 0.33 | 0.20 |
| Mean(b) | -2.44 | -1.81 | -1.33 | -0.81 | -0.22 | 0.15 | 0.53 | 0.94 | 1.62 | 2.29 |
| SD (b) | 0.36 | 0.31 | 0.30 | 0.38 | 0.33 | 0.34 | 0.36 | 0.36 | 0.52 | 0.42 |
| Mean (c) | 0.10 | 0.11 | 0.11 | 0.08 | 0.08 | 0.09 | 0.11 | 0.11 | 0.10 | 0.12 |
| SD (c) | 0.06 | 0.06 | 0.05 | 0.06 | 0.05 | 0.06 | 0.06 | 0.06 | 0.06 | 0.06 |

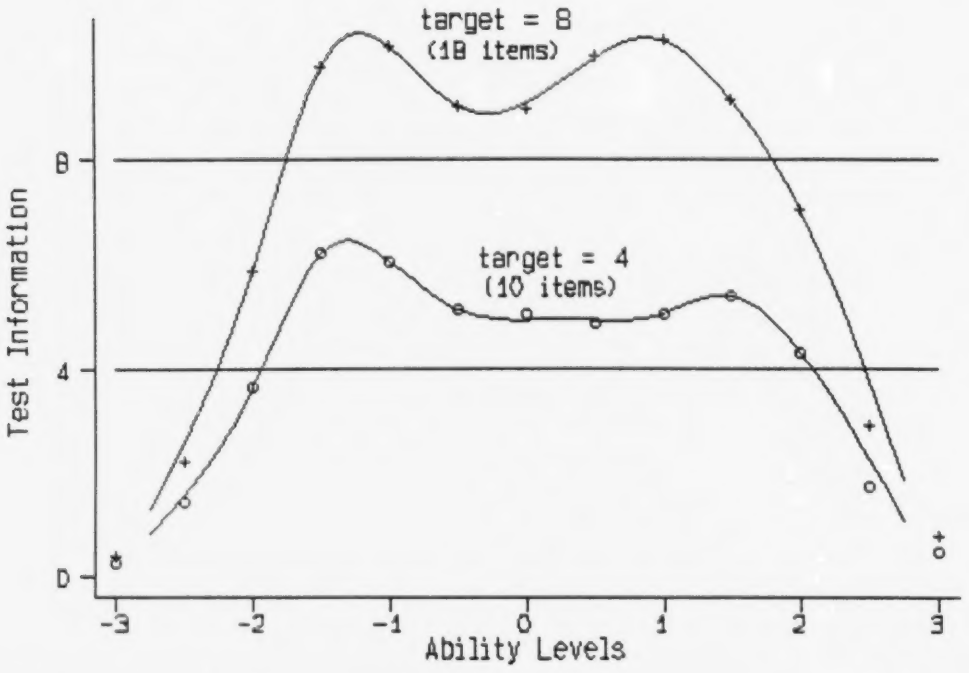


FIGURE 1. Routing test information curves

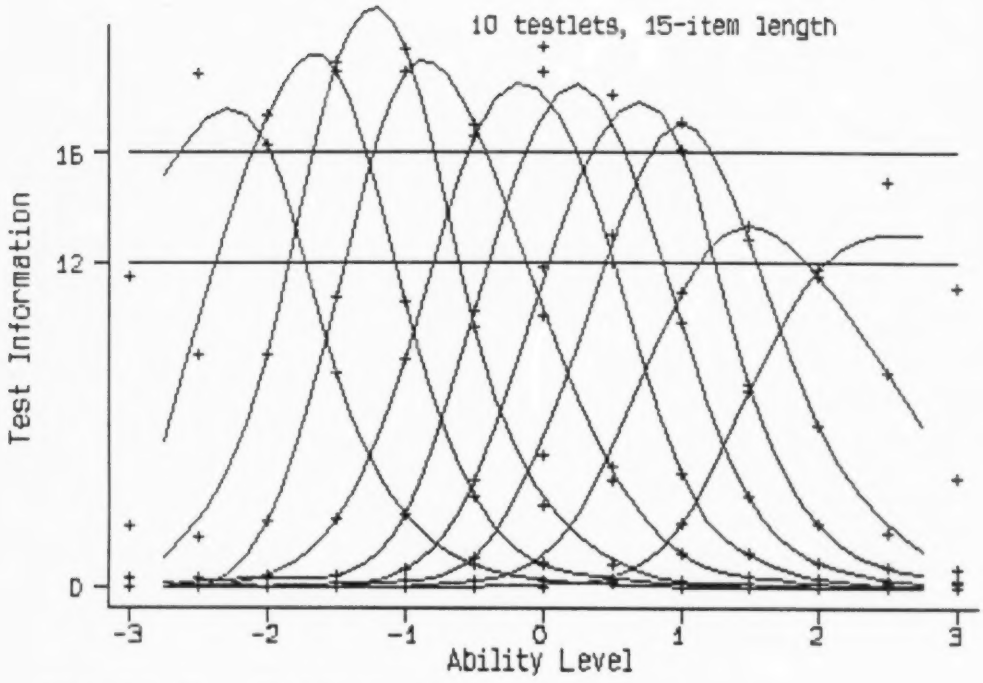


Figure 2. Testlet information curves

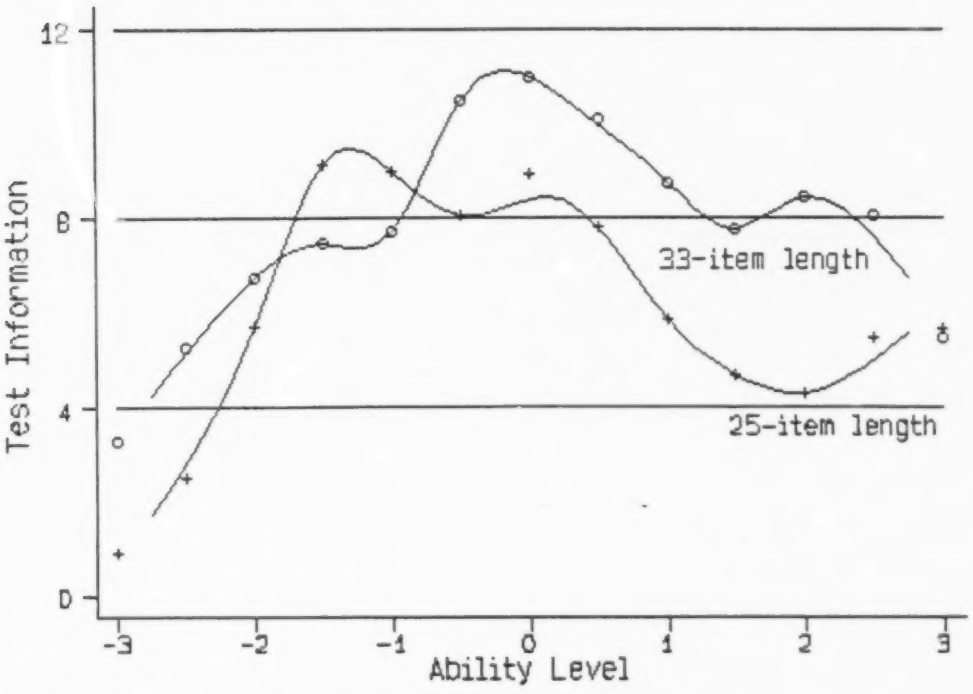


FIGURE 3. Test information curves